Erschienen in: Abel, Andrea/Zanin, Renata (Hrsg.): Korpora in Lehre und Forschung. Bozen-Bolzaro: Freie Universität, 2011, S. 155-178.

Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen[1]

Cyril Belica – Institut für Deutsche Sprache Mannheim

Abstract

Dieser Beitrag betrachtet lexikalisch-semantische Relationen aus einer emergentistischen Perspektive vor dem Hintergrund eines korpusgeleiteten empirisch-linguistischen Ansatzes. Er skizziert, wie eine systematische Erfassung und Auswertung des Kookkurrenzverhaltens[2] von Lexemen – die Analyse der Ähnlichkeit von Kookkurrenzprofilen mit Hilfe von selbstorganisierenden lexikalischen Merkmalskarten und ihre im Diskurs verankerte Interpretation – wichtige Einblicke in die Struktur verschiedenartiger Verwendungsaspekte dieser Lexeme einschließlich ihrer semantischen Nähe ermöglichen. Die vorgestellte Methodik wird dabei – über die explorativ-analytischen Zielsetzungen hinaus – als eine abduktive, auf Theoriebildung zielende Generalisierungsstrategie im postulierten Lexikon-Syntax-Kontinuum verstanden. Zum Schluss werden die Anwendungsmöglichkeiten einiger Komponenten dieser Methodik in der Lexikografie, Lexikologie und Didaktik diskutiert.

1. Das Deutsche Referenzkorpus

In Anlehnung an den Auftrag des IDS, „[…] die deutsche Sprache in ihrem gegenwärtigen Gebrauch zu dokumentieren und zu erforschen […]", hat der Aufbau von elektronischen Korpora am Institut eine lange Tradition. Der in den 60er Jahren des zwanzigsten Jahrhunderts noch neue Gedanke, Analyse und Beschreibung der deutschen Sprache auf computergespeicherte Textkorpora zu stützen, wurde von den Gründungsdirektoren des IDS Paul Grebe und Ulrich Engel in die Tat umgesetzt: Im Jahr 1964 begannen die konzep-

1 Dieser Beitrag bietet einen Rückblick auf die Forschungsarbeit des Programmbereichs Korpuslinguistik am Institut für Deutsche Sprache (IDS) in Mannheim in den letzten Jahren mit einem besonderen Fokus auf die Hypothese der emergenten Natur lexikalisch-semantischer Relationen. Er ist durch eine stilistische Bearbeitung des Transkripts eines am 28. März 2008 in Brixen gehaltenen Vortrags entstanden und bewahrt dessen kursorischen Charakter. Er stellt die Ergebnisse gemeinsamer Arbeit des Autors mit Holger Keibel, Marc Kupietz und Rainer Perkuhn und der Kooperation mit Marie Vachková vor, denen der Autor auch für die Durchsicht und viele Verbesserungen des Manuskripts herzlich dankt.

2 Wegen der klassifikatorischen Vorbelastung des Terminus *Kollokation* in der deutschsprachigen Fachliteratur wird am IDS der Bezeichnung *Kookkurrenz* insbesondere dann der Vorzug gegeben, wenn es darum geht, die theorieneutrale, prä-interpretative *Perspektive* auf die im lokalen lexikalischen Kontext empirischen Sprachmaterials objektiv beobachtbaren (d.h. messbaren) Kohäsionsstrukturen hervorzuheben.

tuellen Arbeiten und die Texterfassung (damals noch auf Lochkarten und -streifen) des *Mannheimer Korpus 1*.
Heute gehört das Deutsche Referenzkorpus des IDS mit seinen ca. 3,4 Milliarden laufenden Textwörtern (Stand: November 2008) weltweit zu den größten und wichtigsten sprachwissenschaftlich motivierten Sammlungen des geschriebenen Deutsch. Es wird im Hinblick auf Größe, Stratifikation, extratextuelle Dokumentation und urheberrechtliche Unbedenklichkeit fortlaufend weiterentwickelt. Wegen der restriktiven Rechtslage ist DeReKo kein offenes Korpusarchiv. Das IDS erwirbt das nicht übertragbare und durch weitere Auflagen eingeschränkte Nutzungsrecht an DeReKo-Textmaterial durch Lizenzverträge mit den (derzeit mehr als 170 einzelnen) Textgebern und räumt seinerseits den DeReKo-Endbenutzern das Recht zur wissenschaftlichen, nichtkommerziellen Nutzung von DeReKo durch individuelle Nutzungsvereinbarungen ein, wie Kupietz und Keibel (2009a) berichten. Sie setzen fort:

> To retain its long-time outstanding reputation as a trustworthy partner of text donors in Germany and abroad, the IDS maintains the highest standards of ethical and meticulous legal conduct pertaining to the DeReKo use and encourages and expects its employees to report any suspected violations of the end user licence agreement.

Vor dem Hintergrund kritischer Überlegungen zum wissenschaftstheoretischen Status empirischen Materials in der linguistischen Forschung (vgl. Keibel / Kupietz, 2009) wird DeReKo seit Anfang der 1990er Jahre nicht als ein „ausgewogenes" oder gar „repräsentatives" Korpus, sondern als eine möglichst umfassende *Ur-Stichprobe* (*primordial sample*) des öffentlichen Schriftsprachgebrauchs konzipiert. Die Komposition von geeigneten auf der DeReKo-Ur-Stichprobe basierenden *virtuellen Korpora* ist bei dieser Strategie aus der Phase des Korpus*aufbaus* herausgelöst und bleibt – unter Einbeziehung der spezifischen Zusammenhänge der konkreten linguistischen Fragestellung und ihrer Implikationen bezüglich der Adäquatheit der zugrundezulegenden Grundgesamtheit – den späteren Korpus*nutzungs*phasen vorbehalten. So wurde zum Beispiel das der aktuellen Version der Kookkurrenzdatenbank *CCDB* (s. Abschnitt 4.1) zugrundeliegende virtuelle Korpus *ccdb-2007* u.a. durch eine per Zufallsziehung vorgenommene quantitative Dämpfung der thematischen Komponenten „Politik" und „Sport" aus DeReKo abgeleitet.

2. Der theoretische Rahmen

Die prinzipielle Nützlichkeit von Sprachkorpora in der linguistischen Forschung und in den benachbarten Anwendungsgebieten ist heutzutage weitestgehend unumstritten. Allerdings wird die Rolle des empirischen Materials in Abhängigkeit von der jeweiligen Zielsetzung, dem gewählten theoretischen Standpunkt, von verschiedenen praxisbezogenen Faktoren und nicht zuletzt vom aufgebrachten Maß an wissenschaftsmethodischer Strenge sehr unterschiedlich verstanden. Der „dokumentarisch-unterstützende" und der „Vermutungen überprüfende" Beweggrund sind wohl die zwei vorherrschenden Motivationen eines typischen DeReKo-Nutzers. Im sprachtheoretischen Kontext wird das Korpus darüber hinaus oft nur als ein – aus der Perspektive einer frei wählbaren Systematik – möglichst erschöpfend *zu beschreibender* Datenkörper gesehen.

Will man jedoch einen korpus*geleiteten* empirisch-linguistischen Ansatz konzipieren, der *explanatorische Theoriebildung* zum Ziel hat, so ist es ratsam, zunächst das einzusetzende Gerüst von Annahmen und Lösungsstrategien darauf zu hinterfragen, ob es mit dem epistemischen Status des Forschungsgegenstandes (d.h. der Sprache) überhaupt vereinbar ist. In ihrer Betrachtung eines genuin korpuslinguistischen Forschungsprogramms, in dessen Kontext auch die im vorliegenden Aufsatz beschriebenen Arbeiten entstanden sind, stellen Kupietz und Keibel (2009b) ihre Auffassung der weit verbreiteten – obgleich ungesicherten – Annahmen und Lösungsstrategien der linguistischen Theoriebildung folgendermaßen dar:

> Vermeiden sollte man zunächst vor allem die folgenden Credos der theoretischen Linguistik – nicht weil sie erwiesenermaßen falsch wären, sondern eher, weil nicht erwiesen ist, dass sie zutreffen: erstens die Annahme, dass Sprache als formales System adäquat fassbar ist; zweitens die Annahme, dass Dekomposition uneingeschränkt als Explanationsprinzip anwendbar ist; drittens, eine vollständige Theorie als Forschungsziel anzustreben; und viertens, die sprachliche Kompetenz zum Gegenstand der Forschung zu machen.

Sie diskutieren unter anderem, wie man bei dem Versuch, die „breite Kluft zwischen der angestrebten theoretischen Beschreibungsebene einerseits und der phänomenologisch zugänglichen Ebene des Sprachgebrauchs andererseits" zu schließen, überhaupt zu hinreichend abgesicherten Erkenntnissen für eine explanatorische Theoriebildung kommen kann und betonen in diesem Zusammenhang „die Notwendigkeit, sich dem Untersuchungsgegenstand Sprache mit möglichst wenigen Vorannahmen über diesen Gegenstand selbst zu nähern". Im Rahmen eines derart konzipierten Forschungsprogramms wird

das Phänomen *Sprache* aus einer emergentistischen Perspektive ausgeleuchtet, der zufolge „alles Regelhafte und Konventionelle in der Sprache ein Epiphänomen des Sprachgebrauchs ist und von den Sprachteilnehmern fortlaufend ausgehandelt wird". Dieses ontologische Prinzip – so die Grundhypothese – durchdringt die Sprache auf allen ihren Ebenen und ist z.B. an der Konventionalisierung syntaktischer Strukturen und an der – besonders im Hinblick auf das Hauptthema dieses Beitrags relevanten – Entstehung und Festigung denotativer und konnotativer Aspekte lexikalisch-semantischer Relationen konstitutiv beteiligt. Seine Wirkung – so die psychologische Arbeitshypothese, die hier nur angedeutet werden soll – manifestiert sich in der Interaktion von zwei miteinander verschränkten kognitiven Mechanismen[3]:

- Im spontan-assoziativen Kontext nimmt der sehr weit gefasste Begriff der *Ähnlichkeit* eine Schlüsselrolle ein. Unsere Fähigkeit, im Alltag auch sehr unkonkrete, verschwommene Ähnlichkeiten unwillkürlich und schnell zu erkennen – besser gesagt, unsere Unfähigkeit, das permanente unwillkürliche Erkennen von Ähnlichkeiten außer Kraft zu setzen – ist lebensnotwendig, da Situationen und Kontexte, in denen wir uns bemühen, uns zurechtzufinden, nie vollkommen identisch sind. Ohne die spontane Sensibilität für Ähnlichkeiten wären frühere Erfahrungen für zukünftiges Verhalten nicht ohne analytisches, schlussfolgerndes Denken nutzbar, sie wären nicht assoziativ generalisierbar. Das gilt insbesondere auch für sprachliche Generalisierungen.
- Im symbolisch-klassifikatorischen Kontext werden gezielt regelhafte Zusammenhänge in der individuellen Erfahrungswelt aufgespürt, typischerweise in Form von abstrakten A-posteriori-Generalisierungen. Im Bereich der Sprache manifestieren sich solche ordnenden Mechanismen schon in der Fähigkeit von Sprachteilnehmern, auf der Basis ihrer Spracherfahrung Ad-hoc-Hypothesen über grammatische Regelhaftigkeiten zu formulieren. Bei Ontogenese, Phylogenese und bei jedem einzelnen Sprachverarbeitungsakt eilen die spontan-assoziativen kognitiven Strategien des „Sich-Zurechtfindens" den symbolisch-klassifikatorischen generell zeitlich voraus.

Eine dementsprechend zentrale Rolle kommt in dem hier vorgestellten methodologischen Ansatz dem Begriff der Ähnlichkeit – „als einem graduellen Gegenkonzept zur *Identität*" – zu (Kupietz / Keibel 2009b: 44). Es wird angenommen, dass man – von normativen bzw. terminologischen Zusammenhängen abgesehen – den vermöge gradueller Ähnlichkeit zwischen Situationen

3 Vgl. die Annahme „[…] human language faculty is instantiated in the mind as an equilibrium between two interdependent general cognitive agencies […]" bei Vachková / Belica (2009).

und Kontexten entstandenen unscharfen Konzeptualisierungen von sprachlichen Phänomenen zwar mit Hilfe von A-posteriori-Klassifikationen eine Pseudo-Identität verleihen kann (um sie symbolisch referenzierbar zu machen), dass aber diesen Explikationen immer eine Verzerrung innewohnt, z.B. in Form von Vereinfachung oder Überzeichnung.[4]

Wenn aber Regelhaftes und Konventionelles wie angenommen im Sprachgebrauch (und gemessen am Erfolg der Kommunikation) ausgehandelt wird, dann ist es auch plausibel anzunehmen, dass die im Diskurs fortlaufend stattfindenden Aushandlungsprozesse konkrete Spuren in Sprachkorpora – d.h. in den Aufzeichnungen von Kommunikationsprozessen im Allgemeinen – hinterlassen, welche mit geeigneten strukturentdeckenden Methoden zu ihren Entstehungsbedingungen zurückverfolgt werden können. Im Hinblick auf die hier diskutierten lexikalisch-semantischen Relationen bedeutet es, dass man die Rekonstruktion dieser Aushandlungsprozesse mit einer „Spurensuche" in zwei verschiedenen „Strata" gleichzeitig vorantreiben sollte: in den *lokalen* lexikalischen Kontexten, in denen die einzelnen lexikalischen Einheiten gebraucht wurden einerseits, und in den *globalen*, situativen Kontexten, in denen jene Aushandlungsprozesse jeweils stattgefunden haben andererseits. Da die zu betrachtenden – lokalen und globalen – Kontexte untereinander im Allgemeinen nie vollkommen identisch sind (s.o.), wird man, um die Überprüfung ihrer Zugehörigkeit zu einem bestimmten Aushandlungsakt und somit um einen Hinweis für ihre fallbezogene Relevanz bestrebt, die Explikation ihrer Identität möglichst meiden bzw. hinauszögern und stattdessen wie oben dargelegt mit dem graduellen Konzept der Ähnlichkeit arbeiten müssen.

Im Folgenden wird gezeigt, wie man diese allgemeinen Überlegungen im Einzelnen mit Hilfe einer großen Anzahl von Kookkurrenzprofilen, die auf Kookkurrenzanalysen höherer Ordnungen basieren, und unter Anwendung des Funktionsprinzips von selbstorganisierenden Karten (Kohonen 1990), einer Art von künstlichen neuronalen Netzen, schrittweise operationalisieren und zu *lexikalischen Merkmalskarten* (*Lexical Feature Maps*) gelangen kann, die u.a. wertvolle Einsichten in die semantische Struktur lexikalischer Einheiten bieten. Ein wesentlicher Unterschied dieser Herangehensweise zu anderen, überwiegend auf *information retrieval* ausgerichteten Ansätzen, wie etwa der *Latent Semantic Analysis* (*LSA*, s. zum Beispiel Deerwester et al. 1990), besteht darin, dass hier die den multidimensionalen semantischen Raum definierenden Einheiten ihrerseits durch komplexe Kookkurrenzstrukturen definiert sind, in denen die in den *lokalen* Kontexten auftretenden Regularitäten und Fluktuatio-

4 Selbstverständlich sind derartige Explikationen ungeachtet aller damit verbundenen Schwierigkeiten oft unentbehrlich und sinnvoll.

nen detaillierter festgehalten werden können als bei der dokument-orientierten Verfahrensweise. Der zweite wichtige Unterschied liegt in der postulierten psychologischen Verankerung dieser Kookkurrenzstrukturen begründet, eine Annahme, die insbesondere bei der Auseinandersetzung mit syntagmatischen Mustern (s. Abschnitt 3.1) als recht naheliegend erscheint: Auf diese Art definierte syntagmatische Muster weisen eine derart hohe Rekurrenzrate in den uns vorliegenden Aufzeichnungen der Kommunikationsprozesse auf, dass es – wenn man einmal das Konzept der graduellen Ähnlichkeit als das konstituierende Prinzip von sprachlichen Generalisierungen akzeptiert hat (s. oben im Abschnitt zu *spontan*-assoziativen kognitiven Mechanismen) – geradezu unverantwortlich erscheint ohne einen triftigen Grund die Vermutung zu verwerfen, dass sie von individuellen Sprachteilnehmern – und dennoch intersubjektiv – mitunter als sprachliche *Gestalt* wahrgenommen und bei Sprachproduktion und -rezeption auch als Ganzes verarbeitet werden. Sie können also insofern als *psychologisch reale* Objekte betrachtet werden, als sie möglicherweise beim Sprachverstehen nicht zwingend immer dekomponiert werden, wiewohl sie im Sinne traditioneller sprachsystemischer Kategorien dekomponierbar sind (und analog bei der Sprachproduktion). Nähere Aufschlüsse dazu und eine auf experimentelle Daten gestützte Argumentation werden von den geplanten interdisziplinären Untersuchungen im Schnittpunkt von Korpuslinguistik, Psycholinguistik und kognitiver Psychologie erwartet.

3. Lokaler Kontext

3.1. Kookkurrenzanalyse

Zur Erfassung von Regularitäten in den lokalen lexikalischen Kontexten in sehr großen Korpora wird am IDS meistens die Analysemethode „Statistische Kollokationsanalyse und -clustering“ (Belica 1995) eingesetzt. Sie implementiert einen erweiterten iterativen Algorithmus zur Extraktion von *Kookkurrenzen höherer Ordnung*, welche als diskontinuierliche n-Tupel (im Unterschied zu *n-Grammen*) von Lexemen mit variierenden relativen Positionen zueinander (anders als *positionsgebundene n-Gramme*) auftreten können. Es werden standardmäßig Kookkurrenzen beliebiger Ordnung ermittelt, wobei das auszuwertende Kontext-Fenster dynamisch bestimmt wird. Die extrahierten Kookkurrenzen werden mit Hilfe eines Clusteringverfahrens in Form von Baumhierarchien angeordnet, welche die zunehmend feinkörnige Kookkurrenzstruktur einzelner Ordnungen visualisieren. Das Analysemodul ist seit 1995 in das IDS-Korpusrecherchesystem COSMAS (*Corpus Search,*

Management and Analysis System) integriert und online aufrufbar. Abbildung 1 zeigt einen kurzen Ausschnitt aus einer Web-basierten Präsentation der Ergebnisse der Kookkurrenzanalyse für das Wort *machen*. Die in der letzten Spalte aufgelisteten syntagmatischen Muster (z.B. „macht sich|mir … große Sorgen") können als stellvertretende Illustrationen der eigentlichen Kookkurrenzen verstanden werden und helfen hauptsächlich, die aufgedeckten Kookkurrenzcluster einerseits zurück auf Sprachdaten, andererseits auf die Intuition eines kompetenten Sprechers zu beziehen. Sie sind oft ein nützlicher Ausgangspunkt für die qualitative Interpretation einzelner Kookkurrenzen (vgl. Perkuhn 2007).

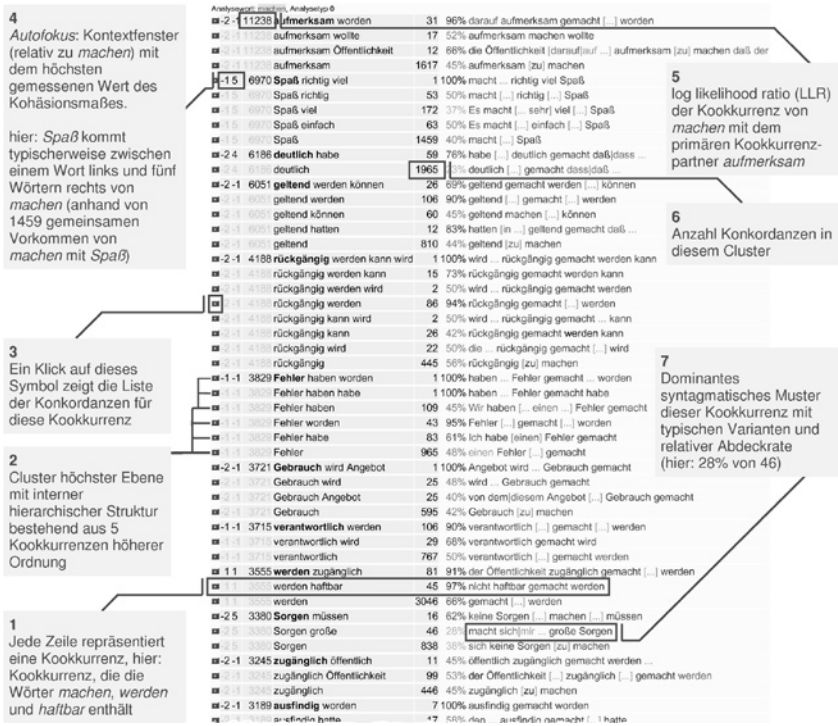

Abb. 1: Annotierte Web-Präsentation der Kookkurrenzanalyse des Wortes machen (Ausschnitt)

Das Analysemodul ist sprachunabhängig, wie im Anhang A anhand der Analyse des Wortes *why* in einem englischsprachigen Korpus illustriert wird.

3.2. Kookkurrenzprofile

Die Gesamtheit aller quantitativen Ergebnisse der Kookkurrenzanalyse zu einem gegebenen Analyseobjekt (einem Lexem, einer Wortverbindung usw.) wird als *Kookkurrenzprofil* des Objektes bezeichnet und stellt – informell gesagt – ein Kondensat seines Gebrauchs dar. Es erfasst sowohl dominante Wortverbindungsstrukturen wie auch subtile Varianzphänomene im lokalen lexikalischen Kontext des analysierten Objektes, und bietet dadurch eine detaillierte Auskunft über die syntagmatische und paradigmatische Einbettung des Objekts im Sprachgebrauch aus präferenzrelationaler[5] Sicht.

4. Globaler Kontext

Die Argumentation zu situativen Kontexten im Abschnitt 2 ließ die Frage offen, wie diese Kontexte, in denen konventionalisierte lexikalisch-semantische Relationen in der Interaktion der Sprachteilnehmer ausgehandelt werden, erfasst, bzw. in den Korpora aufgespürt werden können. Im Folgenden wird gezeigt, dass diese globalen Kontexte sich in der Struktur wechselseitiger Beziehungen einer großen Anzahl von Kookkurrenzprofilen manifestieren, obwohl dies auf den ersten Blick überraschen mag, da Kookkurrenzprofilen nur die lokalen Kontexte zugrunde liegen.

4.1. *CCDB* – das empirisch-linguistische Methodenlabor

Um die wechselseitigen Beziehungen zwischen Kookkurrenzprofilen eingehend untersuchen zu können, wurde im Jahr 2001 am IDS eine „korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs“ (s. Keibel / Belica 2007) begründet. In ihrem Kern enthält sie große Volumina sekundärer empirischer Daten, eine Kookkurrenzdatenbank, die durch mehrfache, unterschiedlich parametrisierte Kookkurrenzanalysen des gesamten Grundformenvokabulars eines großen Korpus[6] errechnet wurde. Derzeit enthält sie mehr als 220.000 nicht-triviale Kookkurrenzprofile pro Parametersatz, bis zu 250 Top-Level-Clusterknoten pro Kookkurrenzprofil und bis zu 100.000 Konkordanzzeilen pro Kookkurrenzcluster. Da diese Analysen mit Hilfe des im Abschnitt 3.1 beschriebenen Analysemoduls durchgeführt wurden, kann prinzipiell jede *CCDB*-Kookkurrenz über das COSMAS-System

5 Vgl. „Generalisierungen über Ähnlichkeitsbeziehungen nennen wir allgemein *Präferenzrelationen*.“ bei Kupietz und Keibel (2009b b: 45); Hervorhebung im Original.

6 Stand November 2008: Dies ist ein ca. 2,2 Milliarden laufende Textwörter umfassendes, auf DeReKo basierendes virtuelles Korpus *ccdb-2007*.

online rekonstruiert und nachvollzogen werden. Die Kookkurrenzdatenbank ist im Sinne eines „gläsernen Labors" teilweise online zugänglich unter <http://corpora.ids-mannheim.de/ccdb/>.

4.2. Paarweiser Vergleich von Kookkurrenzprofilen

Vergleicht man synoptisch die Kookkurrenzprofile zweier Wörter, die intuitiv als „ähnlich" empfunden werden, z.B. die eines synonymischen Paares, so stellt man fest, dass sich auch diese Kookkurrenzprofile teilweise ähneln, sie weisen leicht erfassbare partielle Übereinstimmungen in ihren Kookkurrenzclustern auf. In der Abbildung 2 sieht man beispielsweise, dass die Kookkurrenzprofile von *grinsen* (links) und *lächeln* (rechts) bereits in ihren ersten Zeilen vier Bereiche mit teilweiser Überlappung haben, wobei die Gemeinsamkeiten die Kookkurrenzen mit den Wörtern *sagt*, *verschmitzt*, *freundlich* und *verlegen* betreffen.

Abb. 2: Synoptischer Vergleich der Kookkurrenzprofile von *grinsen* und *lächeln*

Im Umkehrschluss wird angenommen, dass man Objekte, deren Kookkurrenzprofile messbare Überlappungen aufweisen, als *in ihrem Gebrauch ähnlich* bezeichnen kann. Es bleibt eine offene Forschungsfrage, wie ein optimales Ähnlichkeitsmaß für Objekte mit der Komplexität von Kookkurrenzprofilen konzipiert sein soll. Das derzeit in der *CCDB* verwendete Maß beruht auf dem Termfrequenz/Suchwortdichte-Ansatz (*tf-idf*, s. zum Beispiel Salton 1989: 348). Einen ersten Eindruck davon, welches Konzept von Ähnlichkeit dieses Maß operationalisiert, gewinnt man, indem man zu einem gegebenen Wort die Wörter ausgeben lässt, die mit ihren Kookkurrenzprofilen dem Wort gemäß diesem Maß am stärksten ähneln. Als ein Beispiel zeigt Abbildung 3 die Liste der 20 Wörter, deren Kookkurrenzprofile in der *CCDB* dem Kookkurrenzprofil des Wortes *Hindi* am ähnlichsten sind.

1. Chinesisch	8. Arabisch	15. Hebräisch
2. Englisch	9. Italienisch	16. Ungarisch
3. Spanisch	10. Landessprache	17. Amtssprache
4. Türkisch	11. Polnisch	18. Tschechisch
5. Urdu	12. Französisch	19. Russisch
6. Portugiesisch	13. Muttersprache	20. Niederländisch
7. Japanisch	14. Griechisch	

Abb. 3: Wörter mit Kookkurrenzprofilen, die dem von Hindi am ähnlichsten sind

Während die intuitiv leicht nachvollziehbare Ähnlichkeit dieser Wörter – im Sinne von semantischer Nähe – zu *Hindi* unspektakulär ist, da es sich überwiegend um Fälle von Kohyponymie handelt und manche dieser Beziehungen möglicherweise bei Priming-Experimenten signifikant häufig elizitiert würden, ist die Situation bei dem Wort *Charakteristikum* (s. Abbildung 4) spürbar komplexer und jede Vorwegnahme von Ergebnissen experimentalpsychologischer Untersuchungen erscheint hier fragwürdig und unseriös.

1. Merkmal	8. Element	15. Stilmittel
2. Eigenheit	9. Besonderheit	16. Parameter
3. Eigenschaft	10. Charaktereigenschaft	17. Charakter
4. Eigenart	11. Ausformung	18. Eigentümlichkeit
5. Ausprägung	12. Stilelement	19. Ereignis
6. Charakteristik	13. Kriterium	20. Spielart
7. Anliegen	14. Charakterzug	

Abb. 4: Wörter mit Kookkurrenzprofilen, die dem von Charakteristikum am ähnlichsten sind

Im Allgemeinen kann man den folgenden Zusammenhang erwarten: Je facettenreicher die (denotativen und konnotativen) semantischen Strukturen sind, die mit einem *Lexem* (d.h. *nicht* mit einem seiner Denotate) typischerweise assoziiert werden, desto heterogener und divergenter ist die Menge der Lexeme, die ihm *in ihrem Gebrauch* ähnlich sind. Diese Vermutung wird im Folgenden weiter erkundet.

Zusammenfassend soll an dieser Stelle betont werden, dass eine begleitende interdisziplinäre sprachpsychologische Untersuchung von Ähnlichkeitsrelationen und -strukturen zwischen *Lexemen* für eine adäquate korpuslinguistische Modellierung lexikalischer Semantik unverzichtbar ist, und dass sie eine mindestens ebenso große Herausforderung wie die meistens leider allein diskutierte mathematisch-statistische Operationalisierung darstellt. Sobald die Sprache selbst, und nicht z.B. die durch sie transportierte Information empi-

risch untersucht und erschlossen werden soll, muss der Bezug zu den kognitiv-psychologischen Wurzeln von Sprache nicht nur umrissen werden, sondern in der gesamten Forschungsmethodik durchgehend aufrechterhalten bleiben.

4.3. Vergleich von mehreren Kookkurrenzprofilen

Die im Abschnitt 4.2 skizzierten Überlegungen zur partiellen Ähnlichkeit zweier Lexeme (basierend auf partieller Ähnlichkeit ihrer Kookkurrenzprofile) kann man mit dem folgenden Gedankenexperiment fortführen: Wie oben angedeutet, ist es möglich, zu einem gegebenen Lexem λ eine Menge von Lexemen Σ zu nennen, die dem Lexem λ in seinem Gebrauch ähnlich sind. Dabei können diese Lexeme auf sehr unterschiedliche Art zu λ ähnlich sein, d.h. sie können jeweils einen sehr spezifischen Verwendungsaspekt mit λ teilen. Hat man einen sehr großen Vorrat an „verschiedenartig" ähnlichen Lexemen in Σ zur Verfügung, so „berühren" diese das Lexem λ von jeweils unterschiedlichen Seiten, sie bringen jeweils andere Verwendungsaspekte von λ zum Ausdruck. Im Idealfall wäre man in der Lage, zu *jedem* bekannten Verwendungsaspekt des Lexems λ mehrere Lexeme aus der Vorratsliste Σ zu nennen, die eben diesen Verwendungsaspekt mit λ teilen. Wäre es dann zusätzlich möglich, die Lexeme aus Σ so zu Teilmengen zu gruppieren, dass jede Teilmenge einem der sich abzeichnenden Verwendungsaspekte entspricht, so könnte man sagen, dass diese Teilmengen das Lexem λ Aspekt für Aspekt „abgetastet", „kartiert" und dadurch letztendlich bestimmt haben – etwa so, wie man mit geschlossenen Augen durch Abtasten einen Gegenstand zu erfassen versucht.

Etwas formaler ausgedrückt: Über die Analyse der Struktur der Kookkurrenzprofile und des multidimensionalen Vektorraumes, den diese Kookkurrenzprofile vermöge einer Ähnlichkeitsmetrik aufspannen, gelangt man zu Erkenntnissen über die Struktur bestimmender Verwendungsaspekte von lexikalischen Einheiten, insbesondere über die Struktur ihrer lexikalisch-semantischer Relationen. Im Sinne der im Abschnitt 2 zitierten Maxime „[…] sich dem Untersuchungsgegenstand Sprache mit möglichst wenigen Vorannahmen über diesen Gegenstand selbst zu nähern", wird dazu, d.h. zur Analyse bzw. Dimensionsreduktion des Vektorraumes, der Ansatz von neuronalen Netzen – in Form von selbstorganisierenden Merkmalskarten – aufgegriffen. Abbildung 5 zeigt eine mögliche lexikalische Merkmalskarte für das Lexem *Quark*.

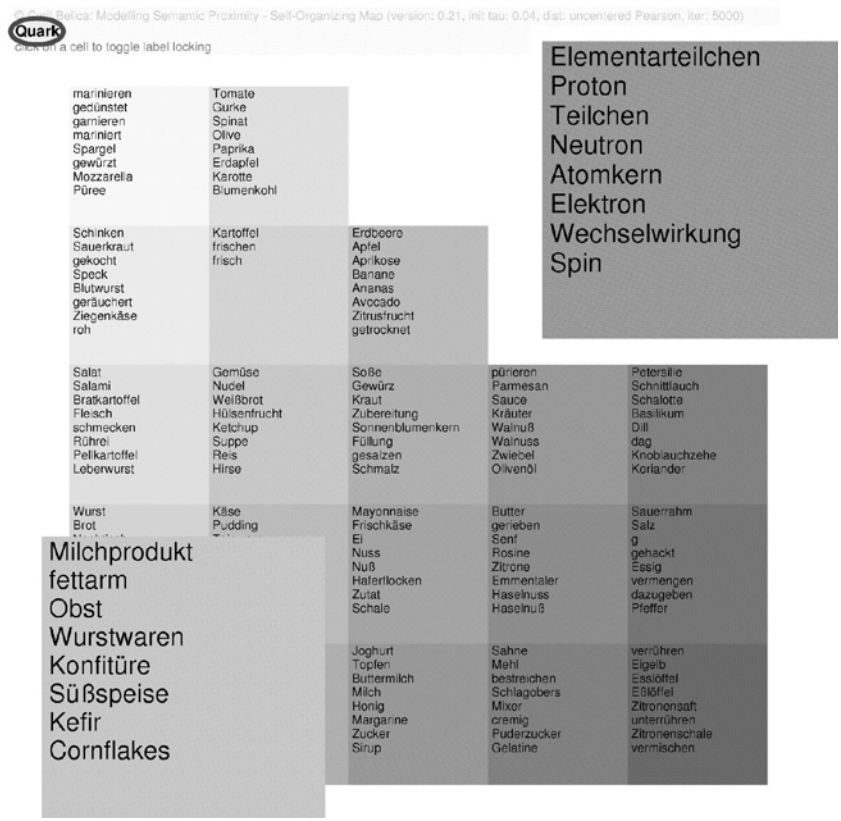

Abb. 5: Selbstorganisierende lexikalische Merkmalskarte[7] für das Lexem *Quark*

Diese Merkmalskarte enthält ausschließlich Lexeme, deren Kookkurrenzprofile dem Kookkurrenzprofil des Lexems *Quark* besonders ähnlich sind. Sie wurden mittels Selbstorganisation so angeordnet, dass die topografische Nähe der Lexeme auf der Fläche möglichst der Ähnlichkeit ihrer Kookkurrenzprofile untereinander entspricht. So sind z.B. die Kookkurrenzprofile von *Konfitüre* und *Süßspeise* nicht nur beide dem Kookkurrenzprofil von *Quark* ähnlich, sondern sie sind – so folgert man – auch untereinander ähnlich, da sie sich im Verlauf der Selbstorganisation nahe beieinander, in demselben Quadrat (s. linke untere Ecke der Merkmalskarte) stabilisiert haben. In der größten Entfernung davon, im Eckquadrat rechts oben, findet man allerdings Lexeme, die

7 Die Vergrößerung der zwei im Text diskutierten Quadrate in dieser Abbildung ist nicht ein Bestandteil der automatischen Visualisierung und wurde zwecks besserer Übersichtlichkeit nachträglich von Hand vorgenommen.

man zwar auch leicht mit dem Lexem *Quark*, kaum aber mit den Lexemen *Konfitüre* oder *Süßspeise* assoziiert, z.B. *Proton*, *Neutron*, *Elektron*. Es liegt daher nahe anzunehmen, dass das Lexem *Quark* in mindestens zwei[8] unterschiedlichen globalen, situativen Kontexten gebraucht wird, die hier verkürzt als *Milchprodukt*-Kontext und *Elementarteilchen*-Kontext bezeichnet werden[9]. In Anlehnung an die Argumentation im Abschnitt 2, wonach man „den vermöge gradueller Ähnlichkeit zwischen Situationen und Kontexten entstandenen unscharfen Konzeptualisierungen von sprachlichen Phänomenen […] mit Hilfe von A-posteriori-Klassifikationen eine Pseudo-Identität verleihen kann (um sie symbolisch referenzierbar zu machen)", würde man z.B. aus lexikografischer Sicht das in diesem trivialen Beispiel „aufgedeckte" Phänomen möglicherweise als Homonymie klassifizieren. Die Maxime, sprachsystembezogene *Generalisierungen* – z.B. die Abbildung von emergenten Phänomenen auf sprachsystemische Kategorien – möglichst erst nachgelagert, *a posteriori* durchzuführen, gehört zu den Grundpfeilern der in diesem Beitrag vorgestellten Methodik.

Im Allgemeinen ist die linguistische Interpretation von selbstorganisierenden lexikalischen Merkmalskarten eine nichttriviale Herausforderung. Vachková und Belica (2009) arbeiten die Methodik einer im Diskurs verankerten semiotischen Interpretation derartiger Merkmalskarten aus und zeigen, dass die emergenten Zeichen intersubjektiv auf das mentale Lexikon bezogen werden können („'re-anchor[ing]' [of] the representamen to a substitute semiotic object in […] mental lexicon."). Sie stellen fest, dass der zu interpretierende Signifikant kein einzelnes Lexem, sondern „a blurred group of topologically close lexemes" ist und charakterisieren dessen semiotisches Objekt als „the quality of the common pair-wise relationship between its constituent lexemes". Daraus folgern sie:

> Since the quality of the underlying relationship – in the sense of 'explication' – is available neither in the corpus nor in the self-organizing map, an attempted semiosis, i.e., rendering the supersign meaningful, requires that its emergent semiotic object be sought in the point of intersection of the individual semiotic objects resulting from their interaction within the mental lexicon due to their connotative potential.

8 Auf einen weiteren globalen Kontext des Lexems *Quark*, nämlich *Unsinn*, kann hier aus Platzgründen nicht näher eingegangen werden.

9 Es ist nur ein Zufall, obgleich ein sehr willkommener, dass die zwei diskutierten Quadrate jeweils auch das Lexem beinhalten, das direkt zur Kennzeichnung der zwei erwähnten globalen Kontexte verwendet werden konnte, nämlich *Milchprodukt* und *Elementarteilchen*.

Die im Anhang B abgebildete, teilweise interpretierte und annotierte lexikalische Merkmalskarte für das Wort *Glas* soll die hier skizzierte Gedankenführung illustrieren.

5. Verwandte Fragestellungen und Anwendungsszenarien

Im Vorangegangenen wurde eine aus mehreren Schritten bestehende Methodik dargestellt. In diesem Abschnitt werden drei Themengebiete skizziert, in denen die Ergebnisse bestimmter Teile dieser Methodik verwendet werden können.

5.1. Synonymie und Plesionymie

Erstellt man eine kombinierte lexikalische Merkmalskarte für ein Paar von semantisch verwandten Lexemen, indem man alle Lexeme, die mindestens einem der beiden Lexeme ähnlich sind, dem Prozess der Selbstorganisation gemeinsam unterzieht, so treten die Gemeinsamkeiten und Unterschiede in der Verwendung einzelner Wörter des untersuchten Wortpaares deutlich hervor. Diese Analysemethode (*Contrasting Near Synonyms*, Belica 2006) kann u.a. zum Aufspüren und Dechiffrieren von feinen Bedeutungsdifferenzen bei Quasisynonymen verwendet werden (vgl. Vachková / Belica, 2009 und Schmidts Analyse von *genau* in Vachková / Schmidt / Belica 2007: 18-20).

Abbildung 6 zeigt eine kombinierte lexikalische Merkmalskarte für das Wortpaar *Einsamkeit/Zweisamkeit*. Die Grauschattierung des Hintergrunds einzelner Quadrate entspricht dem Anteil, in dem die darin aufgelisteten Lexeme einem der kontrastierten Wörter im Gebrauch ähnlich sind, von weiß (entspricht dem globalen Kontext von *Einsamkeit*), über hellgrau (deutet auf globale Kontexte hin, mit denen sowohl *Einsamkeit* wie auch *Zweisamkeit* assoziiert werden könnten) bis dunkel (Situationen, die man typischerweise mit *Zweisamkeit*, nicht aber mit *Einsamkeit* verbindet).[10]

10 Die Wahl der Grauschattierung in dieser Publikation ist drucktechnisch bedingt. Im Original wird zur Visualisierung des Ähnlichkeitsgrades in den einzelnen Quadraten eine unterschiedliche Farbgebung für die Ausgangswörter (gelb bzw. rot) verwendet.

Einsamkeit	Zweisamkeit			
Ehe	Familienleben	Zusammensein	Wonne	innig
Kleinfamilie	Umarmung	Lebensabend	Gegenwelt	Innigkeit
ehelich	Flitterwochen	schönen	nachhängen	Männerfreundschaft
außerehelich	Küssen	sorgenfrei	schön	Zwiegespräch
ehelichen	Verliebte	Kindheit	schrecklich	platonisch
Liebesbeziehung	Flirt	Herzblatt	sonderbar	Liebesszene
Liaison	Liebesnacht	unvergesslich	Kindheitserinnerung	Hassliebe
Ehejahr	schenken	unvergeßlich	wunderbar	Haßliebe
unzertrennlich	Liebesglück	Beschaulichkeit	Freundschaft	Heldentum
entfremdet	Venus	Abwechslung	Erlösung	Sexualität
Liebesspiel		Nachtruhe	Privatheit	unerfüllbar
entfremden		Abgeschiedenheit	immerwährend	Erotik
Beziehungskrise		Geselligkeit	Gemeinsamkeit	Selbstverwirklichung
symbiotisch		Idylle	Ganzheit	Unmittelbarkeit
		Muße	Feindschaft	pubertär
		Behaglichkeit	Seligkeit	voyeuristisch
Erniedrigung	Sinnlosigkeit	Weite	Alleinsein	Zärtlichkeit
Schmerz	Vergänglichkeit	Erhabenheit	Sehnsucht	unerfüllt
Qual	Tragik	Sinnbild	Intimität	Liebe
Entbehrung	Endlichkeit	unendlich	Zartheit	unstillbar
Demütigung	thematisiert	Unendlichkeit	Schwermut	Leidenschaft
Leid	Unmöglichkeit	Gefühlswelt	Empfindsamkeit	Verliebtheit
Mühsal	Fremdsein	Metapher	Geborgenheit	Verlangen
Pein	Dasein	Freiheit	Vergeblichkeit	Zuneigung
Not	Ausweglosigkeit	Verlassenheit	Zerrissenheit	Lebenslust
Depression	Heimatlosigkeit	Langeweile	Melancholie	grenzenlos
Hunger	ausweglos	Trostlosigkeit	Trauer	Suff
Neurose	existentiell	Leere	Enttäuschung	Gier
Alkoholismus	Elend	Tristesse	Auflehnung	Todessehnsucht
Anfall	existenziell	Stille	Gefühl	Eifersucht
Krankheit	Entwurzelung	Düsternis	Empfindung	Geldgier
Drogensucht	Selbstzerstörung	versinken	Verzweiflung	abgründig
Vereinsamung	Orientierungslosigkeit	Hoffnungslosigkeit	Resignation	Wut
Isolation	Erstarrung	Ohnmacht	Verbitterung	Selbstmitleid
Teufelskreis	Vereinzelung	Hilflosigkeit	Aggression	Bitterkeit
Armut	Perspektivlosigkeit	Sprachlosigkeit	Frust	Scham
Verwahrlosung	Erschöpfung	Apathie	breitmachen	Selbsthaß
Verarmung	Überforderung	Frustration	Fassungslosigkeit	Selbsthass
Befindlichkeit	Aussichtslosigkeit	Ratlosigkeit	gewichen	abgrundtief
Verelendung	Desorientierung	Niedergeschlagenheit	umschlagen	Angst

Abb. 6: Lexikalische Merkmalskarte für das Wortpaar *Einsamkeit/Zweisamkeit.*

5.2. Lexikografie und Lexikologie

Ein erstes Anwendungsgebiet, das sich aufgrund des bereits Gesagten anbietet, ist die korpuslinguistisch abgesicherte Lexikografie. Der oben erwähnte emergente Charakter der lexikalisch-semantischen und lexikalisch-syntaktischen Strukturen (vgl. z.B. Bybee 1998: 421 in Anlehnung an Hopper 1987) sowie die fruchtbare Verquickung von Welt- und Sprachwissen (vgl. Adamska-Sałaciak 2006: 47) sind zwei Aspekte, die die moderne Lexikografie besonders betont. Sie sind zugleich als zwei bedeutende Analyseperspektiven bei der Bedeutungserfassung zu nennen: Die oben vorgestellte korpusanalytische Methodik lenkt den Lexikografen in die richtigen Bahnen immer dann,

wenn er in Gefahr läuft, Wörter als feste Entitäten von Form und Bedeutung wahrzunehmen, bzw. diese auf abstrakte Art und Weise unabhängig vom Kontext zu umschreiben (anstatt sie z.B. mit semiotisch aussagekräftigem Wortmaterial zu belegen).

Der kontext- bzw. diskursbedingte Facettenreichtum spielt auch bei der Auswahl von Synonymen oder anderen zum Stichwort äquivalenten Ausdrücken eine eminente Rolle, von der die Qualität der lexikografischen Erfassung eines Wortes (bzw. einer Wortkombination) abhängt: Der Kontrast zwischen der lexikografischen Bearbeitung, die sich nur unmittelbar und hermeneutisch auf Korpusbelege stützt, und dem hier angedeuteten Ansatz, der sich die hier vorgestellte Auffassung der lexikalisch-semantischen Beziehungen zu eigen macht, kann als eine Triebkraft der lexikografischen Arbeit verstanden werden. Als Beispiel kann eine neu zu entwerfende konzeptuelle Auffassung von antonymischen/synonymischen Beziehungen dienen, bei der die skalare Beschaffenheit der Gegensätze diskursgebunden betrachtet wird (vgl. Vachková, 2010). Dies ist eine Herausforderung nicht nur für die Lexikografie, sondern auch für die korpusbasierte Lexikologie.

Hiermit ist bei der Anwendung der korpuslinguistischen Methodik die Frage nach der Vernetzung eines Wortes im Rahmen der paradigmatischen Beziehungen und nach der Eruierung der Wortbedeutung auf einer qualitativ neuen Ebene zu beantworten. Daraus resultiert eine Auffassung von lexikalischen Relationen, die nicht von der a priori verankerten Kategorisierung ausgeht: Das Einbeziehen von Kookkurrenzprofilen in die lexikografische Arbeit bedeutet, einen schnellen und übersichtlichen Zugriff auf eine assoziativ signifikante Diskursmenge als Orientierungsskelett zur Verfügung zu haben.

Die laufende Forschung, die sich mit der linguistischen Interpretation von lexikalischen Merkmalskarten befasst, deutet darauf hin, dass der Prozess, in dem sich das kommunikative Potential der Sprache realisiert, als emergente Struktur visualisierbar ist. Die am Prozess der Selbstorganisation beteiligten Kookkurrenzprofile zeigen sich dabei sowohl bedeutungskonstituierend wie auch bedeutungsrelativierend. Vachková (in Vorbereitung) merkt an:

Die Erfahrung, dass eine diskursbasierte Interpretation von Assoziationsnetzen z.B. die lexikografische Bearbeitung von schwierigen Abstrakta besser steuert als eine abstrakte Bedeutungsumschreibung, erhärtet die Schlussfolgerung von Schwanenflugel: „When abstract and concrete words are placed in highly supportive contexts [sprich: *in Clustern von Kookkurrenzprofilen*], the preactivation of contextual knowledge [sprich: durch *Assoziationen evoziertes Diskurswissen, das an Weltwissen gekoppelt ist*] should override processing [*bzw. priming*] difficulties for abstract words".

Da die der Methodik zugrundeliegende iterative Kookkurrenzanalyse neben binären Wortrelationen auch ganze phrasale Muster erfasst, eröffnen sich außerdem der Lexikografie – auf sehr große Korpora gestützt – nahezu unerschöpfliche Möglichkeiten zur Inventarisierung, Klassifikation und Beschreibung usueller Wortverbindungen – z.B. Phraseologismen, Redewendungen, Sprichwörter, kommunikative Formeln, Funktionsverbgefüge – und der in ihnen vorkommenden (insbesondere paradigmatischen) Variation.

5.3. Didaktik

Ein drittes Themengebiet betrifft nur einen Teil der oben dargestellten Methodik: die syntagmatischen Muster. Aus der in diesem Aufsatz eingenommenen Perspektive auf den sprachtheoretischen Status usueller Wortverbindungen, wie sie sich in den syntagmatischen Mustern (s. Abschnitt 3.1) manifestieren, dass sie nämlich *psychologisch real* sind, liegt es nahe, ihren Stellenwert auch im didaktischen Kontext zu überprüfen. Unter dem Arbeitstitel *simulated late partial immersion* wird am IDS eine weitere Anwendungsmöglichkeit diskutiert mit der Zielvorstellung, ein Kompendium usueller Wortverbindungen für Deutschlernende zu konzipieren. Der Entwurf sieht vor, dass zuerst aus der Kookkurrenzdatenbank *CCDB* solche syntagmatischen Muster extrahiert werden, in denen lediglich Kookkurrenzpartner aus einem gewählten Lernervokabular vorkommen. Diese werden dann redaktionell geprüft, ggf. korrigiert und didaktisch aufbereitet, und wahlweise – auf halbautomatischem Wege – um Beispiele ihrer Verwendung im Korpus ergänzt. Das vielschichtige Zusammenspiel von Faktoren wie Sprachanlass/-situation und Lernniveau könnte je nach Zielpublikum angemessen berücksichtigt werden, da es möglich wäre, das dem Kompendium zugrundeliegende Vokabular in seinem Umfang und in seiner Zusammensetzung praktisch ohne zusätzlichen Aufwand beliebig zu variieren. Aufgrund der großen systematischen (v.a. paradigmatischen) und idiosynkratischen Variation der äußeren Formen dieser Kookkurrenzverbindungen ist es allerdings nicht einfach, für sie eine Art kanonische Grundform einzuführen und ggf. zu operationalisieren, die ihr jeweils eine für didaktische Zwecke hinlänglich fassbare Identität verleihen würde.

Heringer (2007) berichtet von einem gemeinsamen Experiment in diesem Sinne unter Verwendung eines erweiterten Zertifikatsvokabulars mit ca. 2500 hochfrequenten Grundformen. Er nennt diese Wortverbindungen „Valenz-Chunks“ und gibt interessante Beispiele für ihre didaktisch motivierte Aufbereitung. Anhang C listet einige Beispiele für syntagmatische Muster, die in diesem Experiment verwendet wurden, in ihrer ursprünglichen, nicht aufbereiteten Form auf.

6. Zusammenfassung

In diesem Beitrag wurde der Begriff der *semantischen Nähe* aus einer emergentistischen Perspektive vor dem Hintergrund eines korpusgeleiteten empirisch-linguistischen Ansatzes diskutiert. Es wurde eine konsistente und für die *explorative Erforschung von Sprache mit explanatorischem Anspruch* adäquate Methodik vorgestellt, die sich ihrem Untersuchungsgegenstand mit möglichst wenigen Vorannahmen über diesen Gegenstand selbst nähert. Die Methodik wurde in vier Schritten präsentiert:

Auf einer *Erweiterung* des Begriffs der *Kookkurrenz* aufbauend wurden zuerst *Kookkurrenzprofile* als kondensierte Aufzeichnungen einer *psychologisch realen* syntagmatischen und paradigmatischen Einbettung lexikalischer Einheiten im Sprachgebrauch aus präferenzrelationaler Sicht konzeptuell eingeführt. Geleitet von einem weit gefassten kognitiven Prinzip der *Ähnlichkeit* – als einem graduellen Gegenkonzept zur Identität – wurde ein Vektorraum von solchen Kookkurrenzprofilen auf der Basis eines sehr großen Korpus konstruiert. Im zweiten Schritt wurde dargelegt, dass dieser Vektorraum mit Hilfe von geeigneten analytischen Methoden – den *selbstorganisierenden Merkmalskarten* – auf seine Struktur untersucht werden kann. Der dritte Schritt bestand darin zu zeigen, dass die aufgedeckten Strukturen als *emergente Zeichen* mittels einer im Diskurs verankerten *semiotischen Interpretation* zu den globalen Kontexten, in denen sich die Entstehung und Festigung denotativer und konnotativer Aspekte lexikalisch-semantischer Relationen im konkreten Sprachgebrauch ereignet, intersubjektiv zurückverfolgt werden können. Die Forderung, sprachsystembezogene *Generalisierungen* – z.B. die Abbildung von emergenten Phänomenen auf sprachsystemische Kategorien – erst nachgelagert, *a posteriori* durchzuführen, war schließlich Gegenstand des vierten Schritts.

Weitere Fortschritte und eine auf experimentelle Daten gestützte Argumentation werden von den geplanten *interdisziplinären Untersuchungen* im Schnittpunkt von Korpuslinguistik, Psycholinguistik und kognitiver Psychologie erwartet.

Anhang A

Der folgende Ausschnitt aus den Ergebnissen der Kookkurrenzanalyse des Wortes *why* illustriert die Sprachunabhängigkeit der verwendeten Analysemethode. Diese Analyse wurde in einem ad-hoc zusammengestellten Geschäftsenglisch-Korpus mit etwa zwei Millionen laufenden Wörtern durchgeführt.

Abb. 7: Ergebnisse der Kookkurrenzanalyse von *why* (Ausschnitt)

Anhang B

Abb. 8: Teilweise interpretierte und annotierte lexikalische Merkmalskarte zu *Glas*

Anhang C

Dieser Anhang zeigt einige Beispiele für syntagmatische Muster aus der Kookkurrenzdatenbank *CCDB*, in denen als Kookkurrenzpartner ausschließlich Wortformen aus dem erweiterten Zertifikatsvokabular mit ca. 2500 hochfrequenten Grundformen vorkommen, vgl. Abschnitt 5.3.

war [... ein] voller [...] Erfolg
nichts zu tun [...] haben
in letzter Sekunde
kann [es ...] passieren daß
nicht [mehr] bezahlen
Licht ins Dunkel
tiefer in die Tasche [...] greifen
mit leeren [...] Händen da ...
kann nicht [...] schaden
Ich habe [... den] Eindruck dass|daß ...
die Nase [...] vorn [zu] haben
in der kalten [...] Jahreszeit
fehlt [nur] noch
steht [... ganz im] Zeichen des|der ...
ganz besonders
Das gilt [...] auch für die
allerdings [...] noch nicht ...
an allen Ecken und Enden
wird|es noch [... Jahre] dauern bis
lässt ... zu wünschen [...] übrig
schon einmal
gerade mal
paßt [...] gut … auf
eine ... Rolle [...] gespielt
funktioniert [...] nicht [...] mehr
eine Lösung [zu] finden
ganz schön
mit dem Rücken zur Wand
wer weiß
wieder auf freiem [...] Fuß angezeigt
Rede und Antwort [...] stehen
an|An erster [...] Stelle
viel Glück
Glück gehabt
Glück [im] Unglück hatte ...
ebenso [...] wie die ...
und manchmal [...] sogar
ist ... der Meinung
kommt nicht von ungefähr
an den Start [...] gehen
fast überall
in deutscher Sprache
ist [... für … nicht] geeignet
ist [...] typisch für ...
überall auf|in der Welt
das Maß der|aller Dinge
ist [noch] offen
richtet sich in erster Linie gegen|an die
für frischen [...] Wind in die
auf die Art und Weise wie ...
ein Zeichen [... zu] setzen
ein Lied [davon zu] singen
wohl [...] nicht
in die Tasche [...] greifen
in der Nacht [zum|auf] Sonntag
war ein voller [...] Erfolg
sieht [das ...] anders aus
schon mal
dient in erster Linie der ...
die|der leeren [...] Kassen
an die frische [...] Luft
in Verbindung [...] gebracht werden
erst einmal
am runden Tisch
zum Opfer [...] gefallen
leer ausgehen
will [... nichts] wissen
nicht [so] recht
in Frage [...] gestellt
ist noch [...] offen
in den ewigen [...] Frieden heimgeholt
Sprung ins kalte [...] Wasser
Es hat [... viel] Spaß [...] gemacht
unter Dach und Fach [zu] bringen
in Verbindung [zu] setzen
unter Dach [und] Fach [...] gebracht werden
auf Eis [...] gelegt
Rede und Antwort stehen
auf frischer Tat ertappt ...
in|ins Krankenhaus [eingeliefert|gebracht] werden
Es ist [... kein] Zufall daß
Professor für … [an der] Universität
in der Nacht [zum|auf] Freitag
braucht [...] nicht zu
Rechnung [zu] tragen
stieg die Zahl der ...
wird nicht [...] müde zu
Grund zum Feiern
ja oder nein
die Tür [und Tor] geöffnet
eine ganze Menge
in der Nacht [zum|auf] Dienstag
in erster Linie
eine ganz normale
macht [...] Spaß
grünes [...] Licht für ...
ernst [zu] nehmen
nicht ernst [...] genommen werden
Temperaturen von|um|bis ... Grad
eine ... Rolle [...] spielen
ins Leben [...] gerufen
unter Dach und Fach
spielt [... keine|eine ...] Rolle
zur Kasse [...] gebeten werden
Auf|auf den ersten [...] Blick
über weite [...] Strecken
zu wünschen [...] übrig
die Nase [...] vorn
nicht [...] anders als
nicht [...] nur [... sondern] auch
grünes [...] Licht [für ...] gegeben
kann [sich|mir nicht] vorstellen daß ...
noch [...] lange nicht ...
Ich glaube [...] nicht daß|dass ...
einen Beitrag [zur … zu] leisten
Kilometer [von ...] entfernt
mehr oder weniger
nur noch
Erfahrungen [... zu] sammeln
nicht wirklich …
in den eigenen [...] vier [...] Wänden
steht ganz im Zeichen des|der ...
schon [...] lange nicht mehr
wohl [...] auch
sich|mir nicht [...] vorstellen daß
erst einmal
Antworten [auf die ...] Fragen
miteinander [...] verbunden sind
nicht gefallen [...] lassen
legt [...] Wert darauf|auf die
grünes Licht [für ...] geben

Literaturverzeichnis

Adamska-Sałaciak, Arleta (2006): *Meaning and the Bilingual Dictionary*. Polish Studies in English Language and Literature. Hg. Jacek Fisiak. Peter Lang.

Belica, Cyril (1995): *Statistische Kollokationsanalyse und -clustering.* Korpuslinguistische Analysemethode. Institut für Deutsche Sprache, Mannheim.

Bybee, Joan (1998): *The emergent lexicon*. CLS 34: The panels. Chicago Linguistics Society. 421-435.

Deerwester, Scott / Dumais, Susan T. / Furnas, George W. / Landauer, Thomas K. / Harshman, Richard (1990): *Indexing by latent semantic analysis.* In: Journal of the American Society for Information Science, 41(6), 391-407.

Heringer, Hans Jürgen (2007): *Deutsch lernen mit Valenz-Chunks*. In: Zeitschrift für Angewandte Linguistik 47, 2007, 3-16.

Hopper, Paul (1987): *Emergent grammar*. Berkeley Linguistics Conference (BLS), 13:139-157.

Keibel, Holger / Kupietz, Marc (2009): *Approaching grammar: Towards an empirical linguistic research programme.* In: Minegishi, Makoto / Kawaguchi, Yuji (Eds.): Working Papers in Corpus-based Linguistics and Language Education, No. 3 (pp. 61-76). Tokyo: Tokyo University of Foreign Studies (TUFS).

Keibel, Holger / Belica, Cyril (2007): *CCDB: A Corpus-Linguistic Research and Development Workbench.* In: Proceedings of 4th Corpus Linguistics 2007, Birmingham. <http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf>

Keibel, Holger / Kupietz, Marc / Belica, Cyril (2008): *Approaching grammar: Inferring operational constituents of language use from large corpora.* In: Šticha, František / Fried, Mirjam (eds.): Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice, Czech Republic. Prague: ACADEMIA, S. 235-242.

Kupietz, Marc (2008): *DeReKo durchbricht Drei-Milliarden-Grenze*. Sprachreport 2/2008: p. 28 - Mannheim: Institut für Deutsche Sprache

Kupietz, Marc / Keibel, Holger (2009a): *The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research.* In: Minegishi, Makoto / Kawaguchi, Yuji (Eds.): Working Papers in Corpus-based Linguistics and Language Education, No. 3 (pp. 53-59). Tokyo: Tokyo University of Foreign Studies (TUFS).

Kupietz, Marc / Keibel, Holger (2009b): *Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit.* In: Konopka, Marek/Strecker, Bruno (Hrsg.): Deutsche Grammatik - Regeln, Normen, Sprachgebrauch. S. 33-50 - Berlin/New York: de Gruyter, 2009. (Jahrbuch des Instituts für deutsche Sprache 2008)

Kohonen, Teuvo (1990): *The Self-Organizing Map.* In: New Concepts in Computer Science: Proc. Symp. in Honour of Jean-Claude Simon, p. 181-190. Paris, 1990. AFCET.

Perkuhn, Rainer (2007): *Systematic Exploration of Collocation Profiles*. In: Proceedings of 4th Corpus Linguistics 2007, Birmingham. <http://corpus.bham.ac.uk/corplingproceedings07/paper/132_Paper.pdf

Schwanenflugel, Paula J. / Gaviska, David C. (2005): *Psycholinguistic Aspects of Word Meaning.* In: Cruse, Alan D. / Hundsnurscher, Franz / Job, Michael/ Lutzeier, Peter (Hrsg.): *Lexikologie. Lexicology. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen*. 2. Halbband. Handbücher für Sprache und Kommunikation. HSK 21.2 S. 1735-1748. -Berlin, New York: Walter de Gruyter.

Salton, Gerard (1989): *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, 1989.

Vachková, Marie (2010): *Zur Erforschung und Erfassung der diskursgebundenen semantischen Kontraste auf der Grundlage des SOM–Modells.* In: Tématické číslo Germanistica Pragensia XX. AUC Philologica 2. Praha. Karolinum. 193-208. ISBN 978-80-246-1799-2

Vachková, Marie (2007): *Adjektive auf -bar in kontrastiver und korpuslinguistischer Sicht. Eine metalexikographische Betrachtung.* Linguistica Pragensia 2, S. 57-74. Praha.

Vachková, Marie / Belica, Cyril (2009): *Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography*. In: IJGLSA 13, 2 [Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis, Rauch, Irmengard and Seymour, Richard K., (eds.). – Berkeley: IJGLSA/University of California Press], pp. 223-260. ISSN 1087-555.

Vachková, Marie / Marková, Věra / Belica, Cyril (2008): *Korpusbasierte Wortschatzarbeit im Rahmen des fortgeschrittenen Germanistikunterrichts*. Zielsprache Deutsch, 3/2008.

Vachková, Marie / Schmidt, Marek / Belica, Cyril (2007): *Prager Wanderungen durch die Mannheimer Quadrate*. In: Sprachreport Sonderheft/März 2007. Auslandskooperationen des Instituts für Deutsche Sprache. S. 16-21 - Mannheim: 2007.

Internetadressen[11]

Belica, Cyril (2007): *Kookkurrenzdatenbank CCDB - V3. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs*. <http://corpora.ids-mannheim.de/ccdb/>

Belica, Cyril (2006): *Modellierung semantischer Nähe: Kontrastierung von nahen Synonymen.* Korpusanalytische Methode. <http://corpora.ids-mannheim.de/ccdb/>

11 kontrolliert am 15. November 2008

Belica, Cyril (2005): *Modellierung semantischer Nähe: Analyse und topografische Visualisierung von Verwendungsaspekten in Self-Organizing-Maps.* Korpusanalytische Methode. <http://corpora.ids-mannheim.de/ccdb/>
DeReKo (2008): *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Sprache 2008-II*, Institut für Deutsche Sprache, Release vom 18.08.2008. <http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html>
Institut für Deutsche Sprache (1991-2008): *COSMAS I/II (Corpus Search, Management and Analysis System).* <http://www.ids-mannheim.de/cosmas2/>